APPENDIX II

SCIENTIFIC AIDS FOR LITERATURE SEARCHING

by

Alden H. Emery

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150003-9

THE SECRETARY'S OFFICE

Scientific Aids For Literature Searching

The ACS first became interested officially in mechanical aids for searching the literature in 1946. A Committee on Punched Cards was appointed under the chairmanship of Roger Adams to explore potentialities and ascertain whether or not the Society should get into the field. This group found that the methods were being used successfully by various persons and institutions. However, each was concerned with a relatively small part of the whole field of chemistry and there was some doubt that the devices could be extended in scope to cover all that would interest chemists and chemical engineers as a whole. The significant accomplishments justified serious consideration and the doubts on broad applicability provided a proper field for Society study.

The Board of Directors was favorably inclined but did not feel that funds were available for this purpose. However, Dr. Adams was authorized to approach industry for financial support. He was successful and in 1947, J. W. Perry was hired to give full time to this project. At the same time, he was made chairman of the Adams committee which was expanded and given the assignment of studying and developing methods for using punched cards in literature searching.

Deliberations of the committee made it clear that the problems of using hand-sorted punched cards for small files are quite different from problems that would be encountered if punched-card equipment or equivalent automatic devices, were used for searching broad ranges of subject matter. In other words, hand-sorted punched cards are well adapted to relatively small collections in narrow fields of specialization, while automatic equipment operated by punched cards, or by other suitable means, e.g. magnetized tape, offered the possibility of achieving comprehensive searches over a broad range such as that covered by *Chemical Abstracts*.

As one looks back on it now, that doesn't seem to be a very startling conclusion, but one must view it in the light of knowledge five years ago, five years of significant progress in this field. In pursuing the matter further, two different approaches were taken. It seemed likely that individuals could be stimulated to conduct investigations on the use of hand-sorted cards in connection with activities that they would probably undertake anyway. It was also believed that the committee could best serve those members of the ACS interested in hand-sorted cards by preparing a book on that subject.

The first led to many informative papers, most of them presented at national meetings and some of them duplicated for distribution to workers in the field. After formation of the Division of Chemical Literature, it became the outlet for these contributions. The second approach resulted in a book published in September 1951 (see C&EN, Nov. 26, 1951, page 5076) of which over 1500 copies were sold in the first three months. While this was in the planning stage, the Board of Directors expressed the opinion that the book should not be an official ACS publication and released all rights to the individuals responsible.

By the time the funds collected had been nearly all expended, the Board of Directors felt that the potentialities of the methods had been demonstrated and publicized and that further support should come from those who would benefit financially from the development. However, the time was not ripe. Fortunately, continuing support was furnished by the Center for Scientific Aids to Learning established at MIT under a grant from the Carnegie Corp. of New York.

Adaptation of automatic equipment to searching broad fields of interest is a much more ambitious undertaking than use of hand-sorted cards. A survey of the situation revealed that existing punched-card equipment, designed for accounting, could be used only with limited advantage as an aid to literature searching. A recent investigation has led to the conclusion that existing digital computers are of limited usefulness as aids to literature searching. Perhaps it may be generally true that machines designed for computing purposes can not be adapted very well for literature searching. At any rate, such a generality applies to existing devices used for accounting and computing.

Once the inadequacies of existing punched-card machines had been established, Perry and his associates persuaded IBM to design a new set of machines specifically for literature searching purposes. The experimental models of these machines were demonstrated at the Diamond Jubilee Meeting. At present, IBM is conducting further design studies looking to building production prototypes in the near future. The first phases of a campaign to build an ultra-high-speed searching machine based on the electronic techniques used in high-speed computers have been initiated. Preliminary design calculations indicate that such a machine could be expected to search the indexes to five million documents per hour.

These developments in the machine realm have provided a basis for experimenting in a preliminary fashion with indexing methods specifically designed to permit machine searching at a high rate of speed. Last July some preliminary results were demonstrated in Washington

[illegible] scope of the project was enhanced to applying machine indexing methods to a broad range of subject matter. This is being done under government sponsorship.

Before beginning this work, a study of the terminology found in the subject indexes of *Chemical Abstracts* was undertaken. It was decided to set aside the names of compounds and of organisms and to prepare a card file of the remaining terms. The file consists of Keysort cards on which each term is entered, together with its dictionary definition, supplemented wherever possible by notes, gleaned from glossaries and other sources, as to recent changes in meaning of the term. This terminology is being analyzed from various points of view, e.g. whether it refers to a process, a material, an abstract concept, or an attribute. Such analysis is preliminary to establishing a machine indexing system. This is an oversimplified statement. Recently we had the opportunity of reading a privately circulated document which required 26 pages to present the philosophy as introduction to 1½ pages of program.

As already noted, the terminology file does not include the names of compounds or of organisms. The representation of molecular structural formulas is a special problem which requires special attention. The latter has been receiving attention on an international scale since the meeting of the International Union of Pure and Applied Chemistry in London in 1947. During the last week of August 1951, members of the ACS Committee on Scientific Aids for Literature Searching and members of the corresponding commission of the IUPAC met jointly at MIT for informal discussions. Subsequently, the commission of the IUPAC selected the Dyson code on a tentative basis, with the understanding that other systems would be carefully checked to see if they provide ideas useful in improving the Dyson system. A committee of the National Research Council has asked the ACS committee to cooperate in making this study as to possibilities of improving the Dyson system. It is also desirable that the ACS committee cooperate as fully as possible in the development program now being conducted under government auspices.

The part of the ACS in this undertaking has varied during the six years of this program. Originally the sponsor was responsible for financing, more recently the Society's part has been wholly advisory through the Committee on Scientific Aids for Literature Searching, a change in name of the original Committee on Punched Cards necessitated by the broadening of the program. It has been a busy committee as shown by the informative reports which regularly have accompanied the minutes of the quarterly meeting of the Board of Directors. The chairman visualizes even greater activity for some members in the future as they are called on to test coding and indexing systems.

Alden H. Emery

APPENDIX III

TRANSLATION OF "PROBLEM OF MACHINE TECHNIQUE IN SCIENTIFIC INFORMATION"

by

L. I. Gutenmakher

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150003-9

PROBLEM OF MACHINE TECHNIQUE IN SCIENTIFIC INFORMATION

BY L. I. GUTENMAKHER

Vestnik Akademii Nauk SSSR, No. 8, 1952, pp. 46-52

(Original in Battelle Memorial Institute Library)

The problem of the development of machine technique for scientific information, posed by the President of the Academy of Sciences of the USSR, A. N. Nesmeianov, is a very complex problem, and its solution will introduce a fundamental change in the methods of handling scientific-technical information, and will contribute to the better organization of scientific research and more rapid application of scientific achievements in the national economy.

The steady growth of the number of scientific investigations and technical developments has resulted in an avalanche of printed material which completely inundates scientists and practicing engineers.

The total number of printed materials produced by the human race is so large that it can be calculated in the hundreds of millions. At the present rate of research, the increase in amount of material will be such that library holdings will almost double every 10 to 15 years. After 50 to 60 years, library holdings may be expected to have increased 15 to 20 times. Specialists in some fields of science are unable to follow the progress in adjoining fields of science and technology. A whole army of bibliographers is at work to help out scientists to systematize and select bibliographical data and compile bibliographies.

Contemporary practice has required the solution of complex technical and scientific problems in the shortest possible time, taking into consideration all applicable data. Most of a scientist's time is spent in selecting literature and obtaining exhaustive information.

Because of the large amount of information scattered among numerous magazines, books, and symposia, the process of finding the necessary information requires so much time that it is easier to conduct an experiment and to make a calculation than to find a description in the literature.

Attempts at classifying informational material by using different library-classification methods cannot effect any radical solution of the problem.

Classification as a method of organized arrangement of material is characterized by the selection of certain criteria as a basis for dividing information into individual "nonintersecting" groups. For example, bacteria are classified according to their morphology, their pathologic effect, the conditions of their life and growth, etc. Chemical compounds are classified according to their composition, chemical structure, and physical properties, and according to their fields of application. Electronic appliances are classified according to fields of application, power, characteristics, etc.

Classification consists basically in a variation of some type of criterion. It is impossible to use all possible combinations of all criteria.

An attempt to formulate such a comprehensive and subdivided classification, so that it would produce the simplest possible answers, is doomed to failure, too.

The theory of combinations makes it possible to determine easily the "astronomical" numbers which are obtained by computing the possible number of elementary questions in such a system. Such a number, even in the case of very few initial data, will be of the order of 10^{1000}.

In connection with these difficulties, the selection and perusal of literature is the affair of the individual scientist or specialist. Bibliographers and special sections of libraries, ministries, institutes, and universities assist with this work.

For example, let us attempt to calculate the amount of work necessary to satisfy the primary needs of practical work with respect to scientific information.

The Soviet Union has several million engineers, technicians, scientific workers, etc. Let us assume that each of them requires certain scientific information (bibliography in a specific field) only once a year. This would constitute around 3 to 6 million such requests each year, or 10 to 20 thousand requests a day.

Further, we shall assume that specific information is selected from material of which the volume may be represented, for example, by 1000 pages of text. We shall assume also that one person can scrutinize 100 pages of such text thoroughly per day. In this case, each such item of information will require an average of 10 man-days.

Therefore, finding the total information requested would require 100,000 to 200,000 qualified research workers scrutinizing the available material.

Our approach is based, not on the existing situation concerning research work, but on the desirable scale of such. Even if the volume of desired information could be decreased significantly, a great deal of work still would be required to procure the information necessary in scientific research.

The constantly growing scientific level of Soviet specialists requires a technique of procuring information which would ensure them an average of not only one item of information per year, but several.

It may be assumed that, because of insufficient information, a significant part of the efforts and means of scientific institutions is used needlessly in duplicating investigations already performed.

Much time is spent in the selection of information as the first step in any large research project, because, before initiating such scientific research, it is necessary to become familiar with pertinent data from all literary sources. Quoting V. V. Maiakovsky, it can be stated that a scientist selecting scientific-information material investigates "thousands of tons of 'verbal' ore".

The informational material accumulated in libraries represents a tremendous potential wealth which is the more useful the better the scientific-information service is organized.

I. V. Stalin indicates that the mechanization of labor is a force without which it is impossible to maintain the present or new scales of production. This indication is fully applicable to the problem of scientific information.

Mechanization is not limited to a simple increase in the rate of material selection. The search for information can be directed not only to problem A, located together with information on problem B, but it may happen that information on problem B will be required. Therefore, it may be advisable to establish the relation between A and B and to determine the type of relationship.

The possibility of rapid extraction of compiled data on individual problems will lead, in addition, to a series of other benefits, including the decrease and perhaps the elimination of the steadily increasing "bureaucracy" in science. At the present time, specialists even in related fields hardly understand each other. More and more labor is being expended in finding and applying analogies in the processes, phenomena, and structures encountered in the different fields of sciences. The preparation of material for information-bibliographical machines will require generalization of the results of the most diverse research and investigations.

According to Academician A. N. Nesmeianov, it is necessary to establish the possibility of perusing information relative to a given problem by means of a machine, proceeding from a series of independent criteria.

In this case, the content of each individual item must be identified in the research report by a certain number of simple elementary phrases - the sizes, facts, statements, and criteria. Scientific hypotheses, ideas, results of experiments, principles of devices, operation, physical constants, time, location, and other informational data must be presented in a condensed and well-defined form.

In first approximation, it may be assumed that an average scientific paper will contain between 100 and 200 of these sentences.

Analysis of all incoming material and formulation and recording of the elementary sentences can be performed according to rules and instructions established by the authors of the articles themselves or by specially provided personnel. The major part of this work can be performed by the already established Institute for Scientific Information while editing abstracts on work in the different fields of science and engineering.

The selection of accumulated material for a given problem should be performed by machine.

The problems themselves should be formulated as simply as possible, in the form of elementary sentences. A large number of such problems could be posed simultaneously. During perusal of the information obtained, it is necessary to consider simultaneously only information related to all the problems.

The answer (solution) should contain an enumeration of selected items (selected bibliography) in the form of the serial numbers of these items registered in special libraries and should indicate their contents.

The problem of obtaining photostats or originals of the selected items must be solved separately. This problem also may be solved by using mechanized devices (automatic cameras and other devices).

The basic problem of the automatic device (for brevity, we will call this device a machine) is the compilation of bibliography according to a combination of a series of criteria (problems or questions). For a large number of such criteria, the number of possible combinations, practically, is unlimited.

In certain cases, undoubtedly, a combination may be posed which is not reflected in any available reference. A negative answer to such a question is also useful, because it will indicate the newness of the investigation proposed or the development.

Because the selection is performed according to the content of the material, the machine, according to the basic idea, may answer the most diverse questions with any combination of given criteria.

For example, suppose information on the physical constants of molecules is to be selected and arranged. The problems to be solved by the machine will be formulated as follows: to find, on the basis of existing bibliographical data, the identification numbers of articles, etc., in which these constants of certain compounds have values in the range established by the problem posed.

In this case, the number of criteria may be different, ranging from one to the maximum number possible, depending on the stated problem.

However, the possibilities of a machine built according to the indicated principle are much wider. In a series of cases, the person requiring solution of the problem may be interested, not in a bibliography as such, but in analysis of the contents of articles.

For example, suppose that information material on chemical kinetics and on data concerning the mechanisms of chemical reactions is to be selected and arranged in conformance with the stated problem.

The problems confronting machine selection of information in this case may be formulated as follows:

1. To find articles in which slow reactions are discussed, that is, those reactions in which the "pre-exponential" term has a value within certain given limits.

2. To find articles in which it is indicated that a certain reaction proceeds at the definite rate indicated in the problem, and to indicate data concerning the temperature and concentration.

Therefore, it is possible to assign criteria and the numerical value of certain quantities connected with them, and to require compilation of the remaining related information.

At first, it may appear that limitation of the solution by a certain number of criteria without indicating the source of information has no meaning. However, if the tremendous amount of material that can be examined by the machine method is considered, the expediency and effectiveness of such a method becomes apparent.

Often very urgent answers are required concerning the relation of certain values, concerning the coincidence or noncoincidence of a series of signs, etc. For example, in the above-indicated case, it is interesting

to check whether there are contradictory data for corresponding reactions under the very same conditions or under different ones (for example, a reaction in the temperature range from t_1 to t_2 follows the general kinetic equation, but the temperature range from t_2 to t_3 indicates the presence of a more complex chain mechanism).

The machine method makes it possible to extract accumulated data very rapidly and thoroughly, to compare different factors, to analyze data, etc.

However, to realize such a machine system, it is necessary to solve a series of quite complex problems.

1. Creating an Economical, Well-Defined System for Recording Information

The science that produced the above problem has prepared the means for its solution. For an example, it is sufficient to recall the existence of chemical formulas indicating the structure of matter. The theory of dimensions presents the possibility of expressing different physical values by means of a small number of basic values. Thus, to characterize the phenomena investigated in mechanics, it is possible to designate length as L, mass as M, and time as T as basic values, and the dimension of the mechanical values will be expressed as L^2MT^{-2}; velocity as LT^{-2}; density as ML^{-3}; power as L^2MT^{-3}; etc.

For the characteristics of electromagnetic phenomena, a fourth value is added to these basic values - dielectric permeability ϵ or magnetic permeability μ.

Thus, the words "electric field intensity" may be written by the dimension formula: $L^{-\frac{1}{2}} M^{\frac{1}{2}} T^{-1} \epsilon^{-\frac{1}{2}}$. If the words "electromotive force", "potential", "intensity", and "voltage" appear in the text, all of them can be expressed by the formula:

$$L^{-\frac{1}{2}} M^{\frac{1}{2}} T^{-1} \epsilon^{-\frac{1}{2}}.$$

The compilation of a particular dictionary, generalizing the many specialized dictionaries now existing, may enable us to find information in the most unexpected places.

Many examples are known in which "new discoveries" in one field of science have been used for a long time in another. For example, a feedback in the mechanical booster of the regulator of a steam engine has been used to increase the stability of operation for about 80 years, but, for electron-tube amplifiers, it was "discovered again" only in the 30's of the present century, and only 5 years after this for magnetic amplifiers.

The trend toward greater generalization exists in every field of science. Significant achievements have been attained in this direction by Soviet scientists, e.g., in the field of the theory of oscillation. To generalize the available material, experience with the theory of similarity and analogies of physical phenomena may be useful.

The method of analogy is based on V. I. Lenin's postulate: "The harmony of nature is indicated in the 'astonishing analogy' of differential equations related to different fields of science".

Certain mathematical analogies of acoustical, mechanical, hydraulic, and other phenomena are well know. Academician A. H. Krylov indicated that such "analogies between problems of completely different fields, but resulting in similar differential equations, exist in large numbers".

Can it be that a similarity may exist between the calculation of the movement of stars governed by the sun's attraction and their own gravity and the rolling of a boat, or between determination of the so-called secular inequalities in the motion of stars and the rotary vibration of the diesel multicylinder-engine crankshaft when operating a ship propeller or electro-generator? Furthermore, if such a formula and equations were described without words, then it would be impossible to determine which of these problems is being solved, the equations used being exactly the same.

Therefore, the presence of analogy in such various phenomena makes it possible to describe them in the following form:

a) By a system of generalized equations

b) By formulas of dimension of existing values

c) By dimensionless values (criteria of similarity)

d) By a series of elementary sentences, indicating the purpose and results of investigation or development

In many cases, the most exact and laconic formulations, can be obtained by using mathematical formulas. This symbolic, economical form of recording different concepts is changing constantly and is in the process of further development.

For example, the recording of the algebraical equation $x^3 + ax = b$. would have been represented 400 years ago like this:

X cubus + A planum X aequatur B solido.

At the present time, a very efficient method of symbolic recording of very complex logical concept, operations, and conclusions (for example, the algebra of logic) has been developed.

The utilization of the arsenal of mathematical means produces very commendable results. Experience from the theory of similarity and the introduction of dimensionless values (criteria of similarity) for the evaluation of values encountered, in comparison to basic units, also may be of considerable help in solving the problems.

One should remember the great experience in the development of exact and clear formulations of statements in patent practice. As is well known, the formula of an invention must be recorded in the form of a series of separate elementary sentences, in which case each sentence must be short and self-sufficient.

If all of these efforts in different fields are combined, the result will be of great individual scientific value.

Development of the technique of scientific information according to the idea of Academician A. N. Nesmeianov requires the establishment of theoretical basis for generalizing information, which is the next logical stage in the development of the theory of similarity and of the analogy of phenomena. The successful solution to this problem will lead to even more effective utilization of the concept of dialectic materialism concerning the reflection of the unity of nature in the development of science and to strengthening the general bonds and interrelationships between different fields of science.

2. Development of a Rational System of Classification of Material

If the system of recording tells what to look for, the system of classification will show where to find the material.

If the machine method will make it possible to peruse all accumulated material within a certain time, for example, 10 to 20 minutes, then the problem of setting up the classification system will disappear by itself.

However, unfortunately, the amount of information is so great that, even at the highest possible speed of operation, it would be impossible to peruse the available material in 10 to 20 minutes.

Preliminary calculations showed that, if 100,000 abstracts, each of them containing 100 elementary sentences with 10 words each, are treated in one year, this will amount to 100,000,000 words of text per year. Perusing such an amount in 10 minutes would require about one-hundred-thousandth part of a minute for each word.

When material that has accumulated for 10 years will have to be perused, the speed of operation will have to be increased to a millionth of a second, or the time of operation will have to be lengthened.

Undoubtedly, in many important cases, an increase of the time of perusing all available material according to the given criteria will be recommended, in order to avoid omitting any material. In most cases, however, the fields of information could be limited. For example, if information on cutting tools of lathes is required, it would be foolish to peruse the bibliography on electronic oscillators. Therefore, it would be advantageous to establish certain flexible systems of classification to avoid loss of time in the search.

In connection with the above, it would be advisable to develop an initial variant of classification for a certain period of time and then to establish a general method of material classification (method of trial and error).

In machine classification, the system of contacts in the commutator must be constructed in such a manner that its structure can be rearranged easily when the method of classification is changed.

Since the machine method is a high-speed device for research, the classification system must be simple. The structural scheme of classification probably will have the form of a "tree of knowledge", with a different number of "branches" in sections.

When setting up a program of information research, it would be possible to record a considerable number of "addresses" of such "branches" in which valuable material may be expected to be present.

The modern technique of commutation makes it possible to establish, not one, but several, parallel contacts for each division of information.

If, for example, it is known that given information is of interest simultaneously for chemistry, physics, biology, and measuring technique, then the "addresses" of all divisions of the fields of science connected by general interest may be ascribed to this information.

We are not considering the engineering part of this problem. The difficulties here seem to be even greater, but modern engineering is able to cope with them.

The development of such a machine technique of information research is of considerable scientific and engineering importance, because the results obtained could be applied in automatic machines, telemechanics, and communications. The problem presented is very timely. Because of its importance and its character, this program must be handled by the Academy of Sciences of the USSR.

NOT FOR PUBLICATION

STATINTL [redacted]

SYSTEMATIZATION OF TERMINOLOGY

STATINTL [redacted]

There is no need to emphasize to this audience the advantages of making effective use of chemical literature, particularly for planning and conducting research and development. Nor is it necessary to discuss at length the fact that the continuing expansion of recorded chemical knowledge makes its use more difficult. Searching out and identifying pertinent publications has become a problem of concern to our profession as is attested by the large number of papers on searching presented before this Division during recent years.

The simplest type of literature search is directed to a single bit of information, for example, the melting point of some one compound. If this were the only purpose served by chemical literature, the searching problem would cause chemists relatively little difficulty. Individual facts standing alone are, however, much less important in our science than the correlation of chemical knowledge required when planning and conducting research and development. The required correlations may assume a variety of forms, such as textbooks, monographs, literature surveys, research proposals, etc. Preparing such correlations inevitably requires much time and effort to the different types of tasks. One of these is searching out the pertinent information to be correlated. This task can be reduced to a large degree to a set of routine operations, some of which are so repetitious and tedious in nature that they can be accomplished advantageously by machine.

In considering how best to use automatic equipment in locating information to serve the needs of research we must keep in mind certain basic limitations as to what machines can accomplish. Their ability is outstanding in performing well-defined routine operations without fatigue, and at a high rate of speed. Machines, however, are devoid of any power to evaluate or interpret. This means that the usefulness of machines is limited to identifying what documents are of pertinent interest to a given research problem. The interpretation of the significance of subject matter of documents and the correlation of information contained in them requires the vastly superior

ability of the human mind. The identifying and selecting operations performed by machines can expedite the work of human beings by relieving them of the tedium of reviewing large masses of material not pertinent to the problem at hand. Machine searching enables information seekers to devote their time and effort more fully to the interpretation and correlation of pertinent information rather than to inspection and rejection of information of no interest to a given problem. Our goal is to establish effective teamwork between machine performance of routine clerical tasks and human experts interpreting selected pertinent information.

The starting point for developing our machine searching system is recognition of the fact that the analysis of factual information is a necessary prerequisite to its correlation. If we are interested in collecting and eventually correlating information on the use of catalysts in the high pressure synthesis of methanol, then a preliminary analysis making it possible to direct a search to such factors as "catalyst", "high pressure synthesis", and "methanol" is required. The first step in such an analysis is essentially an indexing operation. Someone well acquainted with the subject must inspect the information and decide which of its aspects will be important in selecting and retrieving pertinent papers. Making this indexing step as simple and easy as possible is one of the goals to be kept in mind when devising our indexing procedures.

Once the indexer has performed his task, the headings that he has selected must be recorded so that the machine can perform searching and selecting operations. One way in which this can be done is by punching patterns of holes in IBM cards. Alternately, the index entries may be recorded as patterns of magnetized dots on computer tape or as patterns of opaque and transparent spots on motion picture film.

Regardless of the medium used for recording index entries, the machine searching operation is based on the matching of patterns used to record index entries with corresponding patterns used to define the scope of the search. With appropriately designed equipment, the scanning operation can be performed so rapidly that it is practical to have the machine search the entire index of a file. For example, searching rates of at least thirty million index entries per hour can be attained [1] with electronic digital machines designed for high speed information searching. As a consequence, no advantages are to be gained by alphabetizing index entries. In fact, machine searching is rendered more effective by holding the entries pertaining to a given paper together as a block. This facilitates identification of papers characterized by a combination of entries, such as "catalyst", "high pressure synthesis", and "methanol".

Identification on the basis of a combination of entries is made possible by building into the searching machine a multiplicity of pattern detecting units, e.g., photoelectric cells or comparator circuits. Each detecting unit may be set so as to respond to some one pattern, which in turn corresponds

(1) Electronic Digital Machines for High-Speed Information Searching. P. R. Bagley. SM Thesis, M.I.T., 1951.

to one of the index entries or similar criteria used to define the scope of a search being conducted. A matching condition generates an electrical pulse in the output terminals of the corresponding detecting unit. In the experimental IBM card scanning machine the output terminals are connected to the terminals of a simple plugboard which can be wired so that selection of a card depends on certain relationships between the criteria used to define a search. Thus -- to start with a simple case -- it is a simple matter to wire the plugboard of the experimental IBM scanning machine in such a fashion that a card will be selected only if the photocells detect all of several patterns, which for convenience we may designate by letters, e.g., by A, B, C or D. A selecting operation requiring all of several patterns to be present is said to correspond to the logical product symbolized for example as:

$$A \cdot B \cdot C \cdot D$$

It is equally easy to wire the plugboard so that selection is based on the presence of any one of several factors and such selection is said to correspond to the logical sum, symbolized, for example, as:

$$A + B + C + D$$

A further simple possibility is to base selection on the presence of one pattern and the absence of another. This type of selection is termed a logical difference and symbolized as:

$$A - B$$

Finally, more complicated relationships can be specified in defining the scope of the search. One might specify, for example, that selection is to occur if any one of a group of several patterns be present together with any one of another group. This might be represented symbolically by:

$$(A + B + C) \cdot (D + E + F)$$

Even more complicated relationships between patterns may be specified, such as:

$$(A + B) \cdot (C - D) + (E - F)$$

As our experiments with indexing and searching systems progress, we anticipate the usefulness of these more complex relationships between pattern matching in identifying pertinent documents.

Before these selecting operations can be carried out by machine, it is necessary, as already noted, to record index entries, e.g., by punching cards, so that machine searching operations may be accomplished. The simplest way to do this is to use appropriately selected patterns of holes, or the like, as means for recording letters and numerals. In this approach, the recording patterns would spell out the index entries in exactly the same way that embossed patterns of dots are used to spell out words in Braille. This simple form of encoding can be accomplished automatically by machines already in existence. Depressing a key corresponding to a given letter or

numeral causes the corresponding pattern to be punched in the card. In other words, an operation very similar to typing suffices to render the index entries immediately searchable by machine. Once the cards have been punched, the searching operations may be directed to any word or combination of words punched into the cards. This approach has the unquestionable advantage of simplicity. On the other hand, it suffers from certain disadvantages. One of these, which is closely related to problems encountered in using conventional subject indexes, is the necessity for the operator of the machine to set it so that searching and selecting is based on the right words. One source of difficulty in this connection involves synonyms. Since the machine operates by pattern matching, a search directed to the patterns of holes representing "mercury" will not produce a positive response when inspecting a card punched to represent "quicksilver". Near synonyms and terminology having overlapping areas of meaning would result in similar and perhaps even more vexing limitations on the usefulness of this simple approach, even though it can almost certainly provide considerable aid in rapid screening of papers and reports.

The above-mentioned, simplest form of encoding is also limited as to usefulness by another factor. This limitation becomes apparent if we consider the situation that would result if we punch the names of animals, e.g., "dogs", directly and then a search is required for all animals (not human) susceptible to rabies. Before the machine could be set to conduct this search, we would have to compile a list of all the animals that might be susceptible to the disease. A theoretical alternative -- quite impractical in terms of machine operations -- would be to base the search on a list of all known animals. It would also not be satisfactory to attempt to meet the search requirements by selecting papers on rabies, since this would result in simultaneous selection of all papers on rabies transmitted to humans as well as to other animals, on control measures, and the like. Along with documents of pertinent interest, the information seeker would almost certainly have to cope with an excessive amount of extraneous material.

Improvement in the discriminating power of the machine searching system would be highly advantageous. In one example, this could be accomplished by going over to a different approach in the encoding of individual animals. Instead of using the punched patterns to represent their names as ordinarily spelled, it would be more appropriate to use the patterns to designate other meaningful symbols as illustrated by the following:

Animal	AN
Vertebrate	AN VE
Mammal	AN VE MA
Dog	AN VE MA DO
Cat	AN VE MA CA, etc.

Aside from this type of class inclusion relationship, it may prove advantageous to construct the code so that the whole is denoted by some of the symbols used to encode its parts. An important example is the coding of place names. A simple code for New England states might be worked out as follows:

United States of America	US
New England	USNE
Maine	USNEME
Vermont	USNEVT
New Hampshire	USNENH
Rhode Island	USNERI
Connecticut	USNECT

If the above indicated symbols are used, then the act of encoding any one of these six states, makes both the code for New England (USNE) and for United States of America (US) available as a reference point for defining and conducting a search. It should also be noted that it would be easy for the machine to distinguish between USNE standing alone and in combinations such as USNEMA, as the machine would be able to distinguish, for example, between "cat" as a separate word and the same three letters as found in "catalog" or similar words. It is apparent, of course, that such coding of whole-part relationship may not prove sufficient for certain purposes. The codes given above do not provide a means for discriminating between the states as to such characteristic features as extensive industrialization or status as one of the original 13 English colonies. In our simple example, we might attach additional symbols to indicate these features of the New England states. In this case, some of the codes would be further extended as indicated below:

United States of America	US
New England	USNE
Maine	USNEME
Vermont	USNEVT
New Hampshire	USNENH, OR
Massachusetts	USNEMA, OR IN
Rhode Island	USNERI, OR IN
Connecticut	USNECT, OR IN

When OR designates one of the thirteen original colonies and IN designates an industrial state, in conducting a search it will be possible to condition the machine so that a combination of symbols within a single code are detected. Thus it would be possible to direct the machine to select information relating to all New England states which are industrialized. This would require searching for the combination USNE and IN. If similar codes are worked out for other states in the Union, it would be possible to direct a search to a combination US and IN and select out those items in which at least one of the industrial states of the Union constituted an index entry.

From what has been said, it is perhaps apparent that the purpose of code construction is to render machine searching more effective and efficient. This is accomplished by incorporating in the code symbolism denoting general terms such as "industrial state" or "original English colony" so they can be used as reference points for defining and conducting a search by automatic equipment. It should be noted in this connection that simplification in a coding scheme can be achieved by disregarding certain distinctions which, though perfectly valid or logical, would not be advantageous in the machine searching system. If, for example, in dealing with the New England states we are never concerned with an individual state as such but rather with the region as a whole we might decide to employ the code USNE, OR, IN for the region as well as for any regional subdivision, such as one of the states. Furthermore, inclusion of the symbols OR and IN in the code for New England would depend on whether the aspects so indicated are sufficiently important reference points for conducting selecting operations by machine. If, for example, the historical status of some of the New England states as original English colonies is unlikely to be of interest, then the corresponding symbols should be omitted from our codes for the region and its individual states.

This discussion of the New England states illustrates another point in connection with code construction, namely, that certain characteristics are more readily and easily determined than others. No doubt attaches to which of the New England states were among the 13 original colonies. In contrast, an arbitrary decision may be required as to whether a given state is to be regarded as industrialized. In general those attributes which involve a minimum of arbitrariness are to be preferred when constructing codes.

Enough has been said perhaps to indicate that the effectiveness of a machine indexing system can be greatly increased by devoting care to establishing the most effective possible code for the terminology used for indexing purposes. It would be impossible to over-emphasize the importance of simplicity as a desirable element in the final code. In striving to achieve maximum effectiveness in the simplest possible fashion, one of the most important problems in code construction is the selection of general terms to build into the code. In our example such general terms were "United States of America," "New England," "industrial state" and "status as one of the 13 original English colonies." This type of general term has come to be spoken of for convenience as a "semantic factor."

To keep the codes for indexing terminology as simple as possible only these semantic factors advantageous as reference points for defining and directing searches should be built into the codes for specific terms. The fact that a given semantic factor may be validly related to a given term does not mean that the factor should be so set up in the code. In fact, care must be exerted to avoid including disadvantageous semantic factors.

Pages V-9 to V-10 illustrate how the semantic factoring technique has been developed. The list at the top of page V-9 presents the first dozen terms from our present code dictionary. It will be noted that the same single factor, namely BARA, has been set up as the code for the closely related terms "abrade," "abrasion" and "abrasive." The same pair of semantic factors, namely BASO and MACI have been assigned to "absorber" and "absorption tower."

This illustrates the fact that we have conducted our semantic factoring so as to establish the more important relationships between terms rather than to differentiate between closely related terms. Other devices, which will not be discussed in this paper, are available for making finer distinctions if these should prove important and significant in conducting searches.

Selecting out terms having some one semantic factor in common generates lists, examples of which are shown for the factors TEXI (textile) and TIME (time). The second of these lists has been narrowed down by a second selecting operation directed to the factor MACI (machine or device) in addition to TIME. Not all time devices are used for measuring time. Hence applying the further factor MESU (measure) would permit us to direct a search to another selected list of index entries.

The two final lists on the second sheet show terms characterized respectively by NALA (chemical analysis) and by this factor in combination with PORE (methods). A certain measure of arbitrariness is unavoidable in deciding which of the NALA terms will also be advantageously designated by the factor for method. In making such decisions, the basic question is whether we wish a given specific term, e.g., argentometry to respond when machine searching is directed to NALA or to PORE or to their combination.

Lack of time prevents a detailed discussion of various difficulties that we have had to work out in developing the semantic factoring technique and applying it to develop codes for terms to be used as index entries in machine searching. This approach to code construction is probably not the most advantageous for encoding chemical compounds and the names of living organisms. Subject to these exceptions, we have found that approximately 300 semantic factors sufficed for constructing codes for over 10,000 terms frequently used for indexing and classifying scientific and technical papers. It is perhaps well to emphasize that the effectiveness of semantic factoring in expediting machine searching is due to the fact that the searching operation may be directed to index terms characterized by any one factor or any

combination of factors. This means that a code based on semantically factored terminology makes available as means for defining and conducting a search, not only the specific terms selected as index entries but also their individual factors alone or taken in combination.

In conclusion, it may be interesting to note that lists of terms having some one factor or some combination of factors in common may prove helpful as an aid when using conventional subject indexes. It is not always easy to call to mind the names of devices which have been used, for example, to measure time or to remember what analytical procedures might be looked up in a conventional subject index to locate needed information. We are hopeful that lists of terminology grouped according to semantic factors would prove effective aids in using conventional subject indexes in addition to greatly improving the effectiveness of machine searching operations.

* * * * * *

abaca	TEXI FIBE
ablution	CELA
abortion	BILO DEDA GEGE
abortifacient	BILO DEDA GEGE DOGU
abrade	BARA
abrasion	BARA
abrasive	BARA
absorb	BASO
absorbent cotton	BASO TEXI
absorber	BASO MACI
absorption band	BASO RALI CAPI
absorption tower	BASO MACI
acaricide	DEDA PESI

TEXI

abaca	crash	mat
absorbent cotton	cloth	mercerization
animalize	decatizing	mosquito net
balloon	denier	muslin
beer	drape	nerve
boat	drill	nylon
bobbin	duck	rayon
bone	felt	sanforizing
"Botany"	gunny	silk
broadcloth	herringbone	synthetic fiber
carbonization	huckabuck	tissue
card	kier boil	vat dyes
chintz	knit	vinyon
cotton	lake	woolen
cottonization	linter	wool
		yarn

TIME

calendar	horography	time bomb
chronology	horology	timeclock
chronometer	hour	time exposure
chronoscopy	hour glass	timekeeper
chronothermometer	isochronon	time lock
clepsydra	metronome	time sheet
clock	month	time study
day	pendule	time zone
decade	periodicity	timing
eon	season	watch
era	sun-dial	year
gnomon	time	

TIME & MACI

calendar
chronometer
chronomthermometer
clepsydra
clock
gnomon
hour glass
isochronon
metronome
pendule
sun dial
time bomb
timeclock
time lock
watch

TIME & MACI & MESU

chronometer
chronothermometer
clepsydra
clock
gnomon
hour glass
pendule
sun dial
timeclock
watch

NALA

aleurometer
analytical chemistry
analyze
analysis
argentometry
ashing
assay
azotometer
Babcock test
Baljet reaction
Baudouin test
Benedict's solution
boat
bomb
bomb calorimeter
bromometry
ceriometry
chlorometry
chromatography
colorimetry
conductometric titration
doctor test
elaidin reaction
electroanalysis
end point
eudiometer
examination
exploration
Fehling solution
gasometric analysis
guaiac reaction
Gutzeit test
Halphen reaction
hematimeter
inspection
investigation
iodometric
iodometry
lactometer
magnaflux test
mercurimetry
Molisch reaction
Nessler reaction
nesslerization
oxidimetry
polarograph
polariscope
potash bulb
radioactive indicator
research
Salkowski reaction
sectrometer
spectrometry
standard solution
titrimeter
titration
volumetric analysis

NALA & PORE

argentometry
ashing
assay
bromometry
ceriometry
chlorometry
chromatography
colorimetry
conductometric titration
electroanalysis
gasometric analysis
iodometric
iodometry
mercurimetry
nesslerization
oxidimetry
spectrometry
titration
volumetric analysis

* * * * * * * *

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150003-9

APPENDIX IV

INFORMATION ANALYSIS FOR MACHINE SEARCHING

by

James W. Perry

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150003-9

INFORMATION ANALYSIS FOR MACHINE SEARCHING

JAMES W. PERRY*

The use of automatic equipment for searching and correlating information has been discussed extensively during recent months. This paper dealing essentially with problems and principles represents an attempt to harmonize the views and experiences of a number of persons, particularly Messrs. M. F. Bailey, B. E. Lanham and J. Liebowitz of the U. S. Patent Office; Dr. S. R. Ranganathan of the University of Delhi, India; Dr. W. A. Himwich of Johns Hopkins University; Dr. Jacques Samain of Paris, France; Dr. Vernon D. Tate of M. I. T.; Dr. Charles Bernier of *Chemical Abstracts*, as well as other members of the American Chemical Society's Committee on Scientific Aids to Literature Searching, of which the writer is chairman.

Purpose of Using Machine Methods for Searching and Correlating Information

As far as science and technology are concerned, the impetus toward applying machine methods to searching out and correlating information may be traced to the great expansion in pure and applied research during the last three decades (1, 2, 3).† This expansion has caused a number of important repercussions. Sheer increase in the volume of available information has made the problem of adequate indexing and classifying much more difficult than ever before. As Huntress (4) has pointed out, there now exists very real danger that a new observation or discovery in the laboratory, once made and recorded, may easily become lost — not in the absolute sense, but by disappearing into large accumulations of records and thereby becoming so difficult to find as to be virtually inaccessible. Anyone seriously concerned for the future of scientific research and development must give heed to this situation. This point was emphasized by Bush (1). Yet another factor is the character of at least some research problems as, for example, cancer investigation. Highly complex problems of this sort generate large masses of data whose exhaustive study and complete correlation by an individual would require memorizing such an enormous mass of detail that the task is scarcely within the ability of a single person's memory (5). Another somewhat less obvious factor is the fact that emphasis on research and development in increasingly narrow fields has brought scientific investigation ever nearer to a point of diminishing returns. Summarizing, it may be said that on the basis of the record of science and technology correlating research results presents problems for which solutions must be found, if further progress is not to be hampered.

Previous Documentation Methods — The Starting Point

In any field, an examination of previous methods is a good starting point for considering improvements. If the searching and correlation of information is to be improved by introduction of mechanical aids, then it is advisable to examine previously used documentation methods for the purpose of searching out operations which, because of their repetitive nature, can be reduced to routine and hence might be performed by machines to advantage.

The conventional index may first be considered. Its use is relatively simple and reasonably quick if a specific bit of information say, the melting point of an organic compound of known structure is wanted. If, however, information relating to a broad subject, for example surface active agents, is to be located by using indexes, then a great deal of time must be devoted to scanning many index entries. If the range of interest involves two or more broad headings, the use of surface active agents in textile processing, then the labor involved in tracking down all the pertinent information through an index may become excessive. Much of this labor is repetitive in nature involving, as it does, the scanning of line after line and page after page of index entries.

The situation with regard to use of conventional classification schemes is similar. Thanks to the

* Massachusetts Institute of Technology, Cambridge, Massachusetts.

† *Ed. note:* footnotes follow the text of the article.

Reprinted from AMERICAN DOCUMENTATION Vol. I, No. 3, August, 1950

skill which classification experts have developed, it is quite often possible to look within the appropriate subclass and rather quickly locate desired information. Thus in a file of patents relating to dyestuffs classified according to chemical structure, there may be little difficulty in locating a patent relating to a certain type of diazo dye. In such a file, however, it is no easy task to locate all the dyes disclosed as applicable to cellulose acetate rayon. In such a search, classification on the basis of chemical structure of the dyestuff helps to such a slight degree that it is practically necessary to examine all the patents in the file.

In general, when using a conventional system, difficulties will be encountered if the question to be answered does not require use of one, or at most a few subclasses of the classification system. This type of difficulty is well recognized by the experts who devise classification schemes. Their skill in anticipating searchers' requirements has achieved impressive successes in using conventional classification methods in the field of technology. However, as the volume of information increases, holding material in each subclass to a reasonable volume requires finer and finer subdivision which of necessity is based on finer and finer distinction between the items being classified. The classification system becomes more highly ramified, and more complex. With increasing degrees of complexity there is an inevitable tendency for the proportion of searches out of line with the classification scheme to increase. Similarly entirely too much time of highly skilled persons tends to be devoted to routine examination of large numbers of items most of them not pertinent to the search being performed.

When considering possibilities opened up by the application of machine methods, the accomplishments of conventional indexing and conventional classification cannot be overlooked. Conventional indexing becomes clumsy not because its basic principle of providing leads to information by carefully selected descriptive attributes and concepts is ineffective, but rather because the inspection of a large number of these leads and to an even greater degree establishing correlations among them, is a very time-consuming operation. Conventional classification causes difficulty, not because the grouping of things which have similar attributes is an ineffective means for arriving at desired information, but by reason of the great amount of personal time that must be spent inspecting an excessively large number of individual items in order to find a few which are of interest. The problem is how to set up a mechanized scheme so that designation of items of information by means of properly selected terminology will enable the use of machines to conduct scanning and selecting operations in such a way as to provide a group of items embracing all those of interest in connection with any given question. Or stated somewhat differently, the problem is how to conduct the analysis of information so that machines can be used to scan the results of analysis and collect a reasonably small group of items worthy of attention and study. It is perhaps obvious that there is a close relationship between the analysis of information preparatory to machine searching and the analysis of information in the course of preparing a conventional index. Furthermore, the processes of setting up and running a machine for the selecting of information from a properly analyzed file are strongly suggestive of the processes of establishing a subclass in a classification scheme and placing appropriate items in that subclass.

Although previous discussion has been in terms of searching, it should be noted that searching operations are also essential in establishing correlations.

Some Characteristics of Available Tools

Previous discussion has implied that the machine methods under consideration will employ selecting equipment (6, 7). Although for the purposes of discussion it is not necessary to indicate in detail how a selector would operate in a mechanical sense, it is helpful at this point to consider its principal characteristics.

Basic to any selector system is the interaction of two different machine elements. One is a scanning device and the other some element submitted to the scanning operation. This latter element may assume various physical forms: a file of cards each characterized by a distinctive pattern of punched holes, a reel of motion picture film with each frame having characteristic patterns of transparent and opaque dots or a steel tape with successive sections characterized by patterns of magnetized

spots. This machine element, the file of punched cards, reel of motion picture film, or roll of steel tape would be scanned at a high rate of speed in such a way that the selecting operation can be directed to any one subsection or combination of subsections of the different patterns formed either by holes on successive cards or by transparent dots on the various frames of the film or by magnetized spots on successive sections of steel tape.

Regardless of differences in mechanical details, each item of information before it can be searched must be analyzed by using appropriate terminology. Before discussing the selection of terminology and the most effective methods of effecting its encoding in terms of subsections of patterns of holes, transparent spots, or magnetized dots, it may be useful to consider briefly what operations the machines can perform, that is to say, what types of sorting operations can be conducted. For the purpose of concisely describing these machine operations, assume at the risk of oversimplification that an individual term, an attribute, concept, or the like used for analyzing and designating information, is represented by a capital letter, and that each capital letter is represented by its own distinctive pattern of holes, transparent spots, or magnetized dots. In designating a given item of information on punched cards, film reels, or magnetic tape, an appropriate succession of letters, e.g., A, D, J, Q, X would be entered. These letters would indicate certain attributes or concepts through which the item of information would be approached.

This oversimplified coding scheme greatly facilitates description of searching operations that can be accomplished by various devices already constructed or under design (6, 7). The simplest operation is finding all items characterized by any single capital letter, for example J. The next more complex operation draws together all items characterized by any one of several attributes, as all items characterized by either A or J or D or Q. Another type of operation is the selection of items characterized by certain combinations of attributes and rejecting those in which any one of the combination is absent, to select all items in which attributes A + C + E + J + V are present is an illustration. Another type of search may be directed to items characterized by having one or more attributes present, with the further provision that certain other attributes must be absent, for example A present with B absent — or, symbolically A·B. Finally, is the possibility of conducting searches based on any combination of these different types. The following represent possibilities:

(A or B) + (C or D)
(A or B) + (C − D)
(A + B) or (C + D)
(A or B or C) − (C + D + E)
(A or B) − (C or D)

These possible machine operations have certain implications, all of which are not immediately apparent. Before considering some of them it seems advisable to direct attention to the intellectual problem of devising the system for analyzing information to be used with scanning and selecting machines. Modern automatic scanning and selecting equipment functions quite differently from devices previously in general use in documentation, and for their use it is necessary to work out a new approach to the old problem of analyzing information.

The Relationship of Automatic Equipment to other Operations

It is highly improbable that machines available at present or likely in the foreseeable future will be able to scan printed documents directly. Printed material will be analyzed by a trained analyst and appropriate designations assigned to each item of information. Analysis for machine use as already noted stands in close relationship to indexing as accomplished at the present time. It is vitally necessary to keep the analysis or indexing step immediately in mind when devising systems for applying machine methods to information problems. It would be easy to devise a scheme in which an impossibly severe burden would be placed on the person who must analyze the documents before mechanized searching operations can be performed. Conversely, anything which can be done to ease the task of the document analyst is a step in the right direction, other things being equal.

Another factor in the situation is the necessity for someone, possibly the searcher, to decide how to set the machines to perform the desired operations. The system of mechanized searching and

correlating must be worked out so that the essential searching operations can be conducted with efficiency and precision.

It follows, therefore, that the identical terminology used in analyzing items preparatory to mechanical search must be employed (1) by the person analyzing the document, (2) by the person who decides how to set the machine prior to conducting machine searching and (3) by the machine in encoded form when actually performing the search. There can be no doubt that precise definition and use of terminology is absolutely essential to success in mechanized searching and correlating. However, mere precision of definition is not enough. The terminology must also be selected in such a fashion as to permit maximum speed and convenience in directing searches to broad generic concepts, to subgeneric concepts, to specific terms, or to any combinations. As might be anticipated, the effectiveness of machine methods and the efficiency of conducting the essential operations of analyzing and searching is much improved if appropriate and necessary terminology is grouped into orderly arrays. How this may best be accomplished will next be considered.

The Analysis of Information

Development of an effective scheme for analysis must take into account, first of all, the polydimensional nature of documentary information. This all important point has been emphasized in recent papers (7, 8). An historical event may be analyzed in terms of time, place, persons and organizations involved, and nature of action occurring. The record of a surgical operation is concerned with different types of variables, the patient's symptoms, the pathological conditions thereby deduced, treatment, and the result of treatment. A chemist describes his experiments in terms of the interacting materials, the reaction conditions, the substances produced and accompanying effects, as the evolution of heat.

Use of recorded information is also in the great majority of instances polydimensional in character. An historian usually will be seeking records dealing with the history of some one country during a certain period of time. He is less likely to be interested in all facts recorded about a given country, province or city, regardless of time of occurrence, nor is it likely that he will wish to be informed of all historical events which are recorded as happening during a given time interval. A chemist will rarely be interested in all syntheses carried out under a given set of conditions, e.g., high pressure. He is much more likely to be interested in the effect of a certain set of reaction conditions on individual compounds or groups of compounds.

Merely establishing a multidimensional system for analyzing information is not enough. It is necessary that information be so analyzed that any question which may be posed is either clearly and obviously in harmony with the system of analysis, or at least capable of being brought into harmony with it. At the same time, the system should be made as simple as possible in order to facilitate the analysis.

Designation of the date of an historical event may illustrate a few important considerations. The most widely used chronology in the Western world reckons time from the birth of Christ. The current year under this system is designated as 1950. Mohammedans, Jews and Chinese use different bases of reckoning. It would suffice for purposes of conducting mechanized searching to specify the time in terms of the Christian calendar. If a request for information should specify the time dimension in, say, the Mohammedan calendar, it would be quite easy to convert the specified Mohammedan year to the Christian year and set the machine accordingly.

In generalizing this example, one might compare the Christian, Mohammedan, Jewish, Chinese, etc., calendars with synonyms in a spoken language. It is, in fact, instructive to observe that just as the interpretation of one synonym into another involves an appreciable amount of overlapping and uncertainty, so also the conversion of a given year of the Jewish calendar into the Christian calendar is not completely unambiguous, as the Jewish New Year does not fall on the same day of the month as the Christian New Year. This means that in interpreting one synonym into another, as in interpreting one Jewish year into the Christian time reckoning, it might be necessary to broaden the search somewhat in order to make sure that no items of pertinent information are overlooked. Broadening the

search may well bring to light a number of items not strictly within the scope of the question originally posed. If the number of unwanted items is kept reasonably small no great damage is done, as their elimination by personal inspection would require little time and effort.

Mere designation of the time of an historical event tells nothing concerning location, things and persons involved, or actions. It is instructive to note that the designation of an event in terms of time, place, things, persons and action is analogous to the designation of a point in three dimensional Cartesian coordinates. Just as citing the x-coordinate of a point leaves one free to assign y- and z-coordinates, so the citing of the time of an event leaves one free to designate place, things, persons and action. It is true that mention of 1066 brings to the mind of an Englishman a battle at Hastings, England, between William the Conqueror and the English King, Harold II. But the fact that such a battle did occur at that time and place with contending commanders using certain weapons is a piece of information which can be recorded and coded in terms of the dimensions of time, place, persons, things and actions.

The analogy between the multi-dimensional character of the analysis of information and Cartesian coordinates must not be allowed to become misleading particularly when considering the problem of designating material objects, all of which previous discussion would relate to a single coordinate, the "thing" dimension. In conducting the analysis of information it may not prove convenient to establish a single array of terminology along the "thing" dimension in the same way that numerals are ordinarily used to measure distance along a coordinate axis. Rather, designation of things from different points of view in terms of purpose (means of transportation, power source, weapon), mode of functioning (electrical, internal combustion, atomic fission), materials of construction (steel, copper, plastics), etc. may be wanted. Different points of view constitute in their own right something akin to a set of subdimensions. How elaborate a system of such subdimensions may prove most effective is an important practical problem. In general, increasing the number of subdimensions will make the task of analyzing information preparatory to machine searching more difficult, but at the same time a more elaborate system of subdimensions will provide more paths of approach along which to direct the machine search.

It becomes clear, therefore, that one of the principal problems in designing a system for analyzing information is the selection of the most useful set of subdimensions within the main dimensions of time, position, things, persons and action. Subdimensions in some cases at least are closely akin to parallel descriptions of the same things; they may differ in the sense that a generalization or deduction differs from experimental observations. Thus the progress of a disease may be described in terms of symptoms or more theoretically in terms of the pathological changes in the patient deduced from the symptoms. To approach the patient's record from either point of view by machine searching it is necessary either (1) to record both in a form amenable to machine searching or (2) to interpret one point of view (e.g., pathological changes) in terms of another (e.g., symptoms) in the same way that a given year in the Mohammedan calendar may be expressed in the Christian calendar. If the second approach is used, it is obvious that difficulties and uncertainties may be encountered.

Since the role of theory in science is to provide correlations between observable facts, the analysis and coding of experiments and observations from a theoretical point of view will undoubtedly facilitate correlations with other facts and observations. If the theory used is poorly established and shaky, then correlations attempted on that basis may turn out to be misleading or spurious. In such a situation, analysis and encoding of the observations themselves, in other words the basic data, will probably prove so useful as to be virtually necessary. The degree of confidence which can be placed in theoretical interpretations is therefore an important factor in deciding the design of a system of analysis and coding for machine searching.

The usefulness of generic and subgeneric terms in defining a field of search has already been noted. Such terms must be accorded careful consideration in organizing the system for analyzing documents, i.e., to work out in detail the coding of the various dimensions and subdimensions of the system.

Machines, either available now or capable of being designed, permit a finite number of relationships to be indicated in the coding itself. Such relationships are thereby made directly accessible to machine operations. All metals might be coded by four-letter combinations with M as the first letter, as illustrated by the following examples.

Metals	M---
Alkali Metals	MA--
Lithium	MALI
Sodium	MANA
Potassium	MAKA
Rubidium	MARB
Cesium	MACS
Alkaline Earth Metals	ME--
Calcium	MECA
Strontium	MESR
Barium	MEBA
Radium	MERA

If this scheme were used, information pertaining to all alkali metals could be located by searching for MA-- or any one of the group, for example potassium, by searching for the appropriate code, e.g., MAKA.

This form of coding specifies only certain family relationships and does not take note of various attributes, for example specific gravity. By using the A or B or C or D type of search, it would not be difficult to search out information pertaining to any metal, lithium, sodium, potassium, having a density less than 1.00. The search would be directed to the coded entries, MALI or MANA or MAKA. Here is a simple example of the ability of the machine to synthesize combinations corresponding to generic terms not specifically designated by the coding itself. This very important feature of machine searching in its practical application clearly implies that only those generic and subgeneric terms which are apt to be used frequently in searching should be incorporated in the coding. Otherwise, the effort in coding may not be justified. Here the strong influence of practical considerations makes itself felt once more.

Inclusion of generic and subgeneric terminology in the coding scheme offers a number of additional advantages. In the first place, inclusion of new terms and concepts into a general scheme is facilitated. Thus, to return to the example, the discovery of chemical element number 87, francium, would present no difficulty. Once the elements of Group IA of the periodic table have been defined as "alkali metals," francium will be coded as MAFR. A second advantage of incorporating generic terminology in the fashion indicated is the fact that the generic terms then become immediately accessible without the person analyzing documents having to indicate them each time a specific term is used. Thus the clerical act of encoding the term "sodium" as MANA automatically indicates the fact that sodium is a metal and in addition an alkali metal. A third advantage is the fact that grouping specific items under a subgeneric heading and, in turn, grouping subgeneric headings under a more generic heading, greatly facilitates the task of defining the subgeneric and generic headings themselves. This is a most important advantage, as precision of definition of terminology is a necessity in machine searching.

It is instructive to consider the effectiveness of a coding scheme set up along the general lines outlined above. For purposes of illustrating this point, consider an idealized and oversimplified case in which five different dimensions have been established, each embracing 100 terms used for information analysis; assume furthermore a file of a million information items, each one of which is designated by one term in each of the five dimensions, and the individual terms are used with equal frequency when analyzing each of the million items of information. If a search is directed to any one term on any of the five dimensions, then 10,000 items, one per cent of the total file, will be selected. If interest only in those items having two terms each in a different dimension in common is specified, then out of a million pieces of information in the file, one per cent of one per cent, or a total of 100 items, will be selected. If one additional term in another dimension is specified, only one single piece of information from the whole file can fulfill the search requirements. As already noted, this is an oversimplified example, yet it does serve to illustrate the power of the method which, although making use of a vocabulary of only 500 terms for analyzing information, nevertheless would permit the user to specify and select a single item by using only three guide points in directing the search.

Conclusion

The flexibility of machines now available or capable of being designed is such that it is now possible to conduct the searching and correlation of information in a fashion related to, yet distinctly different from conventional methods of indexing and classifying.

FOOTNOTES

1. Bush, Vannevar. "As We May Think," *Atlantic Monthly* 176, 101–108 (July 1945).
2. Hill, N. C., R. S. Casey and James W. Perry. "Research and Chemical Information," *Chem. Eng. News* 25, 970 (April 7, 1947).
3. Ranganathan, S. R. and James W. Perry. "External Memory and Research." UNESCO/NS/SL/5; *Journal of Documentation* (in press).
4. Huntress, E. H. "Philosophy of the Classification of Chemical Literature," *Ind. Eng. Chem.* 40, 473–476 (March 1948).
5. Toch, Rudolph, Dr., Children's Hospital, Boston, Mass. Private Communication.
6. Perry, James W. "The A.C.S. Punched-Card Committee. An Interim Report," *Chem. Eng. News*, 27, 754–756 (1949). Cf. also *Ibid.* 28, 3789 (1950).
7. Symposium, "New Techniques in Chemical Literature," *Ind. Eng. Chem.* 42, 1456–1468 (August 1950).
8. Wise, Carl S. and James W. Perry. "Multiple Coding and the Rapid Selector," *American Documentation* 1, 76–83 (1950).

APPENDIX V

SYSTEMATIZATION OF TERMINOLOGY

by

STATINTL [redacted]

CONFIDENTIAL

SECURITY INFORMATION

PROPOSED RESEARCH PROGRAM

25X1A5a1

[redacted]

25X1A5a1

CONFIDENTIAL

SECURITY INFORMATION